J Multimodal User Interfaces (2015) 9:353–376
DOI 10.1007/s12193-015-0189-0

ORIGINAL PAPER

The recognition of acted interpersonal stance in police interrogations and the influence of actor proficiency

Merijn Bruijnes[1] · **Rieks op den Akker**[1] · **Sophie Spitters**[1] · **Merijn Sanders**[1] · **Quihua Fu**[1]

Received: 10 January 2015 / Accepted: 2 July 2015 / Published online: 26 July 2015

Abstract This paper reports on judgement studies regarding the perception of interpersonal stances taken by humans playing the role of a suspect in a police interrogation setting. Our project aims at building believable embodied conversational characters to play the role of suspects in a serious game for learning interrogation strategies. The main question we ask is: do human judges agree on the way they perceive the various aspects of stance taking, such as friendliness and dominance? Four types of stances were acted by eight amateur actors. Short recordings were shown in an online survey to subjects who were asked to describe them using a selection of a number of adjectives. Results of this annotation task are reported in this paper. We explain how we computed the inter-rater agreement with Krippendorff's alpha statistics using a set theoretical distance metric. Results show that for some of the stance types observers agreed more than for others. Some actors are better than others, but validity (recognizing the intended stance) and inter-rater agreement do not always go hand in hand. We further investigate the effect the expertise of actors has on the perception of the stance that is acted. We compare the fragments from amateur actors to fragments from professional actors taken from popular TV-shows.

Keywords Interpersonal stance · Embodied conversational agents · Body language · Affect expression · Data reliability

✉ Merijn Bruijnes
m.bruijnes@utwente.nl

[1] Human Media Interaction, University of Twente, 7500AE Enschede, The Netherlands

1 Introduction

We study police interrogations with the aim of building artificial embodied conversational characters. These characters will play the role of a suspect in a tutoring system by means of which police trainees learn to interview witnesses or interrogate suspects. Trainees learn to see how the behaviour of a suspect is related to their own behaviour. Interpersonal stance is a core construct in the theory that is used to understand and explain how suspects behave in a police interview. Currently, in training sessions, actors play the role of a suspect within a specific scenario based on historical material.

People quickly form impressions of each other's personality and interpersonal stance (attitude). This also holds when people encounter virtual humans [10]. Research has also shown that when several judges were asked to encode interpersonal dispositions of people the agreement was low [20]. Nevertheless, if we build realistic and believable virtual suspect characters we need to pay attention to the relation between observable nonverbal behaviours and the way the police trainee perceives and judges the character and attitude of the virtual suspect. Moreover these characters have to be interpretable in the sense that "the user must be able to interpret their responses to situations, including their dynamic cognitive and emotional state, using the same verbal and nonverbal cues that people use to understand one another" [24]. A virtual human does not have the same cognitive capabilities as a human. This makes an interaction between them what Baranyi and Csapó [6] call an *inter-cognitive communication*. The challenge here is that a virtual human has to give the human the impression it has cognitive capabilities that are similar to that of the human. When successful their interaction should be called an *intra-cognitive communication*.

Basically there are three different methods to follow when building models for the generation of the dynamic behaviours

of virtual humans. In the artist method virtual characters are created and their behaviours and expressions are generated based on intuition after which we observe how the character is perceived. Another term that is used in the literature for the artist method is puppeteering [36]. Contrary to the artist method are the analytical approaches towards finding general rules or statistics about the typical behaviours that express a certain stance. In the design method different designers are given a virtual human and a set of basic behaviours and expressions. For a number of stances, the designers are asked to generate behaviours and expressions with the virtual human that they believe express the given stances. The results are analysed to see how often designers used each of the basic behaviours for each of the stances. The statistical behavioural model is used to generate the most likely (combination and sequences of) behaviours when the virtual characters takes a certain stance. This method was followed by Chollet et al. [12]. In this paper we report on a second analytical method which is based on the analyses of stances played by human actors. The scene of play is that of a face-to-face police interview where one police officer interrogates a suspect. In a number of judgements surveys recordings of human actors were presented to human judges who were asked to label the stance expressed by the actors in the fragments of the interviews. The method raises the following three issues.

(A) The collection and selection of the audio/video fragments that show the behaviours. Do we use actors and how do we generate specific stance behaviours in lab settings, or do we use real-life recordings?
(B) The task of the human judges that label the data. What is the annotation procedure, the label set, is it categorical or continuous?
(C) The way reliability of the labelled data is measured. Can we assume ground truth? How to compute inter-rater agreement?

(A) We used two different types of audio/video recordings: for the first experiment we had non-professional actors play a specified stance in a given interrogation scenario or we asked them to respond to a stance taken by the interviewer in a given scenario. For a second experiment we choose a number of fragments from TV series, showing professional actors play the role of a suspect. (B) We report about different ways of labelling stance: (1) a semi-free annotation format where judges could choose from a given set of adjectives that describe the stance shown. (2) A three-dimensional continuous annotation schema: three 5-point Likert scales for dominance, affiliation and spontaneity. (C) Several methods have been used to measure the validity and reliability of the labelled data. Can we assume ground truth about the stance that actors portray?

With the analysis of acted social, emotional, or stance behaviour of human actors comes the consideration of how natural this acted behaviour is. The question is, can an actor show behaviour on demand of the researcher and how close to real-life behaviour is this on-demand behaviour. Bänzinger and Scherer [5] distinguish three categories of 'naturalness' of recorded behaviours: (1) Natural behaviour occurs in real-life settings and is not directly influenced or controlled by the researcher. (2) Behaviour can be induced in a controlled (laboratory) setting that is designed to elicit the behaviour in which the researcher is interested. (3) Portrayals of behaviour by actors upon instruction by the researchers. It is common for elements from these categories to co-occur, for example induce emotions with the instruction of the to be portrayed behaviour. The underlying feelings and emotions of 'natural' behaviour cannot be directly assessed, they can only be inferred from observations or post-hoc reports by the 'actor'. This is also a problem when using TV clips from professional actors as we have done in this study. These clips have the advantage of showing more natural behaviour than that of amateur actors, but the intent of the behaviour cannot be validated. The underlying emotional 'intent' is available from induced behaviour or portrayals of behaviour as it is part of the instruction or design of the setting that induced the behaviour: 'if you ask for a smile, you get a smile'. In this study we have therefore also asked (less experienced) actors to act out behaviours based on the researcher's instructions. By knowing the actual intent of the acted behaviour, we can investigate whether others perceive the behaviour as intended. Obtaining variances within behaviour or multiple instances of a behaviour from the same actor can be difficult or impossible in 'induced' or 'natural' behaviour. When asking actors to portray behaviour it is possible to get all variables of the behaviour from each actor. This allows comparison of the same behaviour from different actors. In addition, Busso and Narayanan [9] took a "deeper look at the current settings used in the recording of the existing corpora [which] reveals that a key problem may not be the use of actors itself, but the ad-hoc elicitation method used in the recording." They suggest that portrayals will be as close to natural behaviour as possible if care is taken to: (1) contextualize the (social) setting properly. (2) Combine the acting styles 'scripted' and 'improvisation' to have both the influence the researcher needs and the freedom the actor needs in their emotional expression. (3) Give actors the time to prepare or rehearse their acts, or use skilled actors if possible. (4) Define the references used to describe the emotional and social acts as they are often blurred and partial to subjectivity [5].

An actor might know which behaviour will be perceived as the intended behaviour and show this behaviour. This is troublesome for those behaviours that are often misclassified by observers. For example, Strömwall et al. [42] show that there is a difference between what people believe to be indicative

for deception and what is actually indicative. For this reason, an actor who is instructed to show deceptive behaviour can unknowingly show unnatural behaviour; behaviour that a deceiving person would not show. And an observer might rate this behaviour as deceptive behaviour.

Acted portrayals have been challenged as unsuited for applied research purposes. Taking a different stance Bänziger et al. [4] argue that: (a) portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selected real-life contexts; (b) acted portrayals as opposed to induced or real-life sampled emotional expressions display the most expressive variability and therefore constitute excellent material for the systematic study of nonverbal communication of emotions. "In everyday life", they argue, "emotional expressions are directed to receivers with different degrees of intentionality. Some expressions might be truly 'spontaneous', not directed or intentionally regulated to have an impact on a receiver; whereas acted portrayals are by definition produced intentionally and directed to a receiver." In our second experiment we asked judges to score the spontaneity of the actor's stance. This paper is an expansion of earlier work described in [41].

In Sect. 2 we will define stance and we will discuss stance taking in the interesting context of the police interview in which both parties are often not on the same wavelength. In Sect. 3 we will discuss related work. In Sect. 4 we will explain the method of our study and in Sect. 5 the outcomes. In Sect. 6 we will investigate the effect the expertise of the actor has on the perception of his behaviour. We draw will conclusions in Sect. 7.

2 Interpersonal stance

Interpersonal stance (or attitude) refers to "those spontaneous or strategically employed affective styles that colour interpersonal exchanges" [39]. Compared to personality, attitudes are subject to a greater degree of variation over time. Interpersonal attitudes are essentially an individual's conscious or unconscious judgement of how they feel about and relate to another person while interacting with them. Argyle [1] and Leary [27] identify two fundamental dimensions of interpersonal attitudes that can account for a great variety of nonverbal behaviour: affiliation (ranging from positive to hostile) and power (from dominant to submissive).

In training conversational skills, in particular in training to interview suspects, police trainees in the Netherlands use T. Leary's theory of interpersonal relations as a framework for analysing their behaviour towards the interviewee and how the suspect's behaviour is understandable as a reaction to their own behaviour and their style of interviewing, for example, is it more understanding or more offensive. Leary's model, the interpersonal circumplex also known as *Leary's Rose* [27] is used in training conversational skills in various professions as for example in health care [43]. The model is presented by a circular ordering of eight categories of interpersonal behaviour, situated in a two-dimensional space spanned by two orthogonal axis. The horizontal axis affiliation (positive versus hostile) the vertical one is the power axis (dominant versus submissive) (Fig. 1a). The theory says that for the power axis; dominant behaviour is met with submissive behaviour and submissive behaviour is met with dominant behaviour. For the affiliation axis; positive behaviour is met with more positive behaviour and hostile behaviour is met with more hostile behaviour (Fig. 1b). For example, if someone is dominant and positive towards you (DP, leading or helping), you would gladly cooperate with them which is a submissive and positive (SP) stance. People are capable of employing this mechanism, changing their own stance in the hope that the other will follow by changing their stance according to the theory. One can see why the police has interest in teaching such a theory. They deal with uncooperative suspects and have the task to make them more cooperative, Leary's theory provides a clear strategy to attempt this change in stance of the suspect. The main aim of the investigative interview is truth-finding. An important

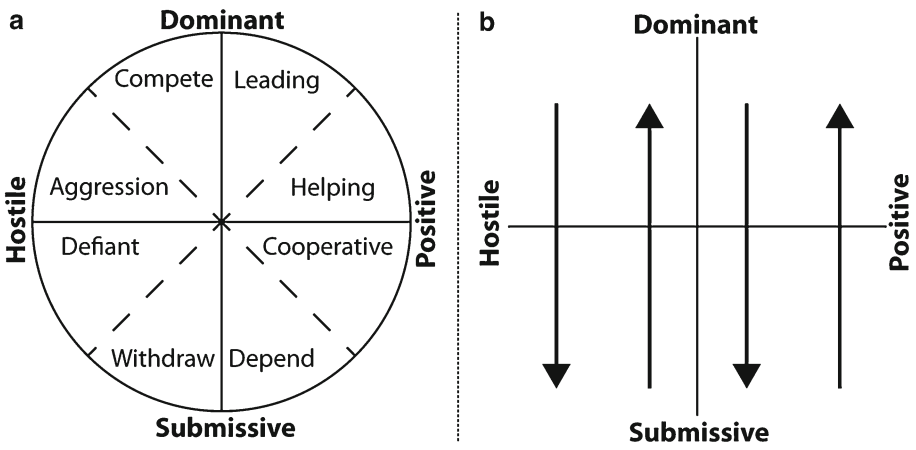

Fig. 1 **a** Leary's Rose with the different stance segments (affiliation is the *horizontal axis*, dominance is the *vertical axis*). We distinguish the segments dominant–positive (DP), submissive–positive (SP), submissive–hostile (SH) and dominant–hostile (DH). **b** The relation between behaviour of the interactors. Dominant behaviour is met with submissive behaviour and vice versa. Positive behaviour is met with positive behaviour and hostile behaviour is met with hostile behaviour

secondary aim is to establish a working relationship with the suspect, rapport building.

Inbau et al. [23] list five essential principles that must be followed by the interviewer in order to decrease the probability of making erroneous inferences from a suspect's behaviour.

1. There are no unique behaviours associated with truthfulness or deception.
2. Evaluate the consistency between all three channels of communication: verbal, paralinguistic (among which response timing and length) and nonverbal.
3. Evaluate paralinguistic and nonverbal behaviours in context with the subject's verbal message.
4. Evaluate the preponderance of behaviour occurring throughout the interview.
5. Establish the suspect's normal behavioural patterns.

Deception and lying happen in police interviews more than in many other encounters and interviews. Suspects make statements to make the other believe something that is not true. A lie is not a special type of sentence with it own linguistic or para-linguistic identifiers. Similarly, stances that are taken deliberately to make some impression on the other are not distinguishable from "sincerely" taken stances. A "strategically employed" stance does not have special types of observable features. This means that we will not expect that the relation between stances and their observable behavioural features is different in the context of police interviews, where people often purposely take a stance, from that in other contexts where people act more spontaneously.

Research has demonstrated the value of the circumplex model for integrating a broad range of psychological topics. Researchers have mostly used Wiggins' English Interpersonal Adjective Scales (IAS) from [49]. Rouckhout and Schacht [38] found a circumplex structure underlying a comprehensive set of Dutch interpersonal adjectives. The configuration was divided into eight segments isomorphic to the IAS octants. Fifteen adjectives from each segment were used to form eight preliminary Dutch interpersonal scales. Table 1 shows the Dutch adjective scales (translated into English) from [38]. A representative random selection of the Dutch adjectives, highlighted in bold, was used in our study.

The technical notion of "nonverbal behaviour" refers to a range of observable aspects of communicative phenomena in human encounters. It includes turn-taking, the switching of speaker and listener roles, prosodic aspects of speech, body postures and facial expressions. All these aspects may reflect stance taking. Here we report on the study of the typical postures and facial expressions that signal the stances. Op den Akker et al. [35] reported on the relation between stance taking and turn-taking in police interviews.

3 Related work

A number of projects aim at the development of virtual characters (ECAs) for social skills training and serious games. For quite some years the Virtual Humans project at the Institute for Creative Technologies (ICT) has been building embodied agents that are integrated into training environments for learning interpersonal skills [24]. Virtual humans have been developed for training negotiations between a commander and a citizen in a law enforcement setting [44]. Olsen [34] describes a system for training police interviewing by means of interaction with a computer simulated suspect. Also, Luciew et al. [29] build virtual learning systems to train police officers in interviewing children who have been victims of sexual abuse and to train police officers to interrogate suspects on that matter. The deLearyous project [48] also used Leary's Circumplex for annotating interpersonal stance. Based on transcriptions of conversations between a human and a virtual character playing the role of an employee in a negotiation between a manager and an employee, their system was able to automatically annotate text on stance and respond appropriately [45,46]. The aim of the EU TARDIS project[1] is to develop an ECA that acts as a virtual recruiter to train youngsters to improve their social skills. In that context [13] considers the analysis of sequences of nonverbal behaviours and expressions. Op den Akker et al. [35] report on the results of annotating stance of human role playing suspects and police officers in a training setting. They found that annotating stance is a hard task and that it is difficult to get satisfying inter-rater agreement when rating on an utterance level, yet that it is possible to get predictable patterns using a majority vote annotation. Behaviours have different meanings depending on the place in a temporal sequence of behaviours. Novielli and Gentile [33] investigate the recognition of the interpersonal stance of the user when having a dialogue with an ECA used as an interface agent. Chollet et al. [12] show recognition of the stance taken by the interviewer so that the virtual human can respond to it.

Birdwhistell [8] studied how people communicate with posture, gesture, stance, and movement. He argued that all movements of the body have meaning (i.e. are not accidental), and that these nonverbal forms of language (or para-language) have a grammar that can be analysed in similar terms to spoken language. He estimated that "no more than 30–35 % of the social meaning of a conversation or an interaction is carried by the words". Communication of interpersonal attitudes is one of the five primary functions of nonverbal behaviour [1]. Mehrabian [31] found that body orientation affects the conversation. Body language comes in clusters of signals and postures, depending on the internal emotions and mental states. Recognizing a whole cluster is

[1] http://www.tardis-project.eu.

Table 1 (Translated from) Dutch adjectives scales for the categories of the interpersonal circumplex (the highlighted words represent the selection used in this experiment)

DH		SH		SP		DP	
Compete	Aggression	Defiant	Withdrawn	Depend	Cooperative	Helping	Leading
Cocky	Unmerciful	**Depending**	Shy	Unpretentious	Affable	Laid-back	Unfaltering
Cynical	Sneaky	Awkward	Submissive	Simple	Sweet	Charming	Extrovert
Dodgy	Pretentious	Authority negligent	Bashful	Maternal	**Tolerant**	Communicative	Garrulous
Offensive	Cruel	Tactless	Quiet	Good	Loving	Cheerful	**Powerful**
Bold	Foul	Impersonal	Ashamed	**Dead serious**	Accommodating	**Spontaneous**	Enthusiastic
Authoritarian	**Biased**	Curt	Reserved	**Humble**	Meek	Excited	Sophisticated
Dominant	Sly	Antisocial	**Doubting**	Innocent	Merciful	Gallant	**Stubborn**
Pugnacious	**Cheeky**	Egocentric	Timid	Overanxious	Composed	Attentive	Proud
Impulsive	**Suspicious**	Loveless	Introvert	Flexible	Willing	Courteous	Resolute
Impetuous	Rebellious	Insincere	**Reserved**	Soft-hearted	**Unprejudiced**	Social	Ambitious
Unabashedly	Cunning	Deceitful	Closed	Casual	Modest	Friendly	Talkative
Thick-skinned	Polished	**Irreverent**	Shy	**Discrete**	Tender-hearted	**Loyal**	Firm
Not timid	Crafty	**Arrogant**	**Artificial**	Timid	**Gentle**	Tactful	Passionate
Not shy	Despotic	Envious	Aloof	Docile	Pliable	**Humane**	Resolved
Unabashedly	Shameless	Intolerant	Naive	Submissive	Sensitive	Helpful	**Lively**

From: [38]

thus far more reliable than trying to interpret individual elements. Smith-Hanen [40] reports a study in the perception of empathy through body postures. The author concludes that more attention should be focused on the nonverbal channels of communication in the training of counsellors. Dael et al. [14] adopted the Body Action and Posture (BAP) coding system to examine the types and patterns of body movement that were employed by 10 professional actors to portray a set of 12 emotions. The authors investigated to what extent these expression patterns support explicit or implicit predictions from emotion theories. The study revealed that several patterns of body movement occur systematically in portrayals of specific emotions, allowing emotion differentiation.

3.1 Nonverbal behaviour perception studies

Physical posture and emotion have been studied in the literature using two similar techniques. The first method involves the participant viewing videotaped actors performing certain actions and the second method involves having the participant sit in a certain posture and then self-reporting their emotions. The use of actors in studying emotions and stance has a number of advantages over collecting real data. But, are acted stances representative of real stance? Or, does the experimental setting in which stance is generated and the fact that subjects know that the stances are acted reduce the validity of the outcomes? As far as we know there has not been a study of the validity of acted *stances*. In the context of *emotion* research in speech the issue has been considered by [50]. Based on a perception experiment, they concluded that acted emotions (especially negative ones) were perceived more strongly than the real emotions. The suggestion is that actors do not feel the acted emotion, and may engage in overacting, which casts doubt on the usefulness of actors as a way to study real emotions. Acting has a particularly strong effect on the spoken realization of emotions. Busso and Narayanan [9] argue that if certain conditions are satisfied professional actors can be used for valid emotion research. Conditions for the "generation of emotions" are that professional well-instructed actors should be used. Enough context should be given to actors for eliciting the emotional state. Asking actors to read a sentence aloud in a "sad voice" is not good practice for building an emotion database. Other conditions concern the perception and descriptions of the emotions. An interesting approach—given the goal of our own project—is the SAL approach in which emotions are induced in a context of a human interacting with a virtual character/machine ([17]).

Many studies have shown some nonverbal behaviour to be important or occurring more during a stance; for an overview see Table 2. Culture influences the stance people take towards others and the meaning of certain behaviours and expressions [3,21]. Endrass and André [19] integrated cultural factors into models of virtual characters. Police interview studies

Table 2 Some typical stance behaviours from the literature

	Dominant	Submissive	Hostile	Friendly
Head movement	Tilt head up, orient head toward other, shake head [11]			Tilt head up, orient head toward other [11]
Hand gesture	Movements directed away [32], high gesture rate while talking [11], initiate hand shaking [11]	Movements directed inward [32], object-adaptor, self-touch [37]	Self-protection gestures [47], folding arms [21,37]	Touch other [37], object-adaptors [37], initiate hand shaking [11]
Posture	Space filling and asymmetric postures [21,32], erected posture [11]	Shrinking postures [21]	Distant postures, barier postures [21]	Physically close postures, close interaction or direct orientation [21,37]
Leg movement	Wide stance of the legs [32]		Rhythmically moving legs [47]	
Facial expression	Facial anger [18], self-assured expression [37], expressing face [37]	Facial sadness [18,37]	Facial disgust [18], facial anger [18,37]	Smile [10,18,21,22]
Gaze behavior	More gaze [10,37], gaze for a long time [37]	avert gaze [37]	gaze for a long time [37]	Mutual gaze [10,37]
Focus of attention		Pay attention to other [37]		Pay attention to other [37]
Turn taking	Overlapping speech [47]	Pause often [32]	Overlapping speech [47]	
Vocalization	Loud voice [32], high pitch [47], high rhythm [47]	Low voice [32]		

have shown how differences between high culture and low culture have impact on how sensitive suspects are for the different interrogation strategies and stances that police officers apply [7]. For a recent survey on affective body expression perception and recognition refer to [25].

4 Method: generating and annotating stances

The method followed in this research consists of two parts. First of all, clips of the interpersonal stances were generated using actors. This will be discussed in Sect. 4.1. The validity of these depicted stances was assessed by means of annotation. The annotation process will be discussed in Sect. 4.2. The results of this annotation process will be discussed in Sect. 5.

4.1 Generating interpersonal stances

The clips of the interpersonal stances were generated by using actors. Eight actors took part in this experiment. Four of them were members of a theatre club and thus, had some acting experience. Each actor had to depict four stances. The stances correspond to the quarter segments of the rose in Fig. 1. The four segments are abbreviated as dominant–positive (DP), submissive–positive (SP), submissive–hostile (SH), and dominant–hostile (DH).

All actors were given the same scenario. They had to imagine they were suspected of shoplifting and in the middle of an interrogation. Then, they watched a computer screen where a video fragment shows a police interrogator addressing them and asking them what happened. So, the only thing the interrogator says is: "Vertel eens, wat is er gebeurd?" ("Tell me, what happened?"). The actors were then asked to give a short response (max. 10 s) with a certain stance. This process was repeated until all four stances were depicted. This produced 32 video recordings. These were used in the survey.

However, the actors differed in the instructions they got on how to depict interpersonal stances. Half of the actors were selected to the 'theory condition' and the other half to the 'role play condition'. Subjects with and without theatre experience were evenly distributed over both conditions. The actors in the 'theory condition' were given theoretical instructions about Leary's Rose [27]. To help them get an even more concrete idea of what the stances mean, several adjectives that capture the meaning of the stances were given. Using these instructions, the actors had to react to the virtual interrogator according to these stances. The adjectives were a random selection from each category of adjectives used in [38], see Table 1. The summary of their instructions were captured in an image, that is shown in Fig. 2. In this condition we tried to ensure that the actors were influenced as little as possible on how they should react to the interrogator on the screen, because it was important that the reaction should be an interpretation of the stances that came from the actors and not from the researchers. The actors in the 'role play condition' were given a specific scenario for each stance that was directly linked to the interrogation setting and to the question of the interrogator. An example of such a scenario is given in Fig. 3. The scenario was supposed to provoke a reaction in a certain stance in a more natural way than was the case in the 'theory condition', as the workload of process-

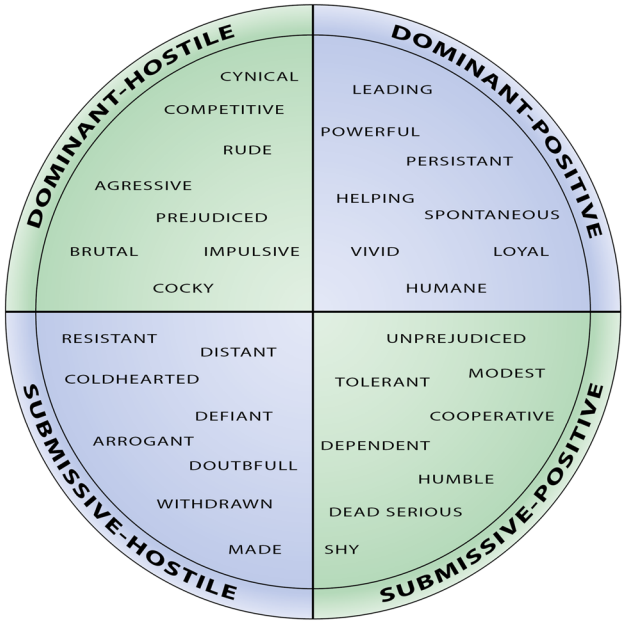

Fig. 2 Summary of the instructions given to the actors in the 'theory condition' in order to express interpersonal stances

Dominant-samenwerkend:
Het verhoor is al even aan de gang en het valt je heel erg mee. De agent is best wel aardig! Daarnaast heb je eigenlijk ook niets te verliezen, want de waarheid komt toch altijd boven water, daar geloof jij sterk in. Je gaat het gesprek dus positief en zelfverzekerd in en zal de agent wel vertellen wat er allemaal is gebeurd. Zo zal het verhoor snel en soepel verlopen!

Fig. 3 Example of a scenario given to the actors in the 'role play condition' to help them express interpersonal stances. This is the scenario for the stance DP. Translation: ***Dominant–positive:*** *The interview is well under way and it isn't that bad. The officer is actually nice! Besides that, you don't have anything to lose because the truth will come out eventually, you strongly believe in that. You approach the conversation positive and with confidence and you will explain the officer what has happened. That way the interview will be quick and smooth!*

ing and interpreting the theory was reduced and actors could put their resources into entering into the part they were playing. Both conditions were taken into account to see whether actors indeed need the information in a scenario format to put down a good performance or if a theoretical instruction is good enough.

4.2 Annotating interpersonal stances

In an online survey judges ($n = 84$) were shown video fragments in which an actor displayed a specific intended stance. The judges had to select a number of adjectives from 32 different adjectives that best fitted how they would describe the stance taken by the actor in the video fragment. A convenience sample was used that consisted largely of students. The participants were each asked to annotate 8 fragments from a total of 64, 32 with sound and the same 32 fragments without sound. The distinction between with and without audio was used to check whether people were better at recognizing interpersonal stances when they also heard what was said and how it was voiced. The video fragments were assigned randomly to the participants, but in such a way that a participant viewed exactly one clip of each actor.

A semi-forced format was used for annotating the fragments, meaning that participants were given a list of 32 adjectives and were free to select any number of adjectives (with a minimum of four) that they thought fit the stances expressed in the fragments. The list of adjectives was the same as used in the theoretical instruction for the actors. Furthermore, the list was presented in a random order to prevent that the first category would be over represented in the data because of order effects. A screenshot of the survey is shown in Fig. 4.

In a forced choice format for emotion recognition subjects had to select one label or word describing the emotion that was shown. The format was debated and some authors advocated a free choice format where subjects were free to choose their own wording to describe the observed emotion. Limbrecht-Ecklundt et al. [28] discussed the pros and cons of both formats. Our format was semi-forced. This overcomes the problems with forced choice. It has the disadvantage that it is less straightforward to compute accuracy and inter-rater agreement.

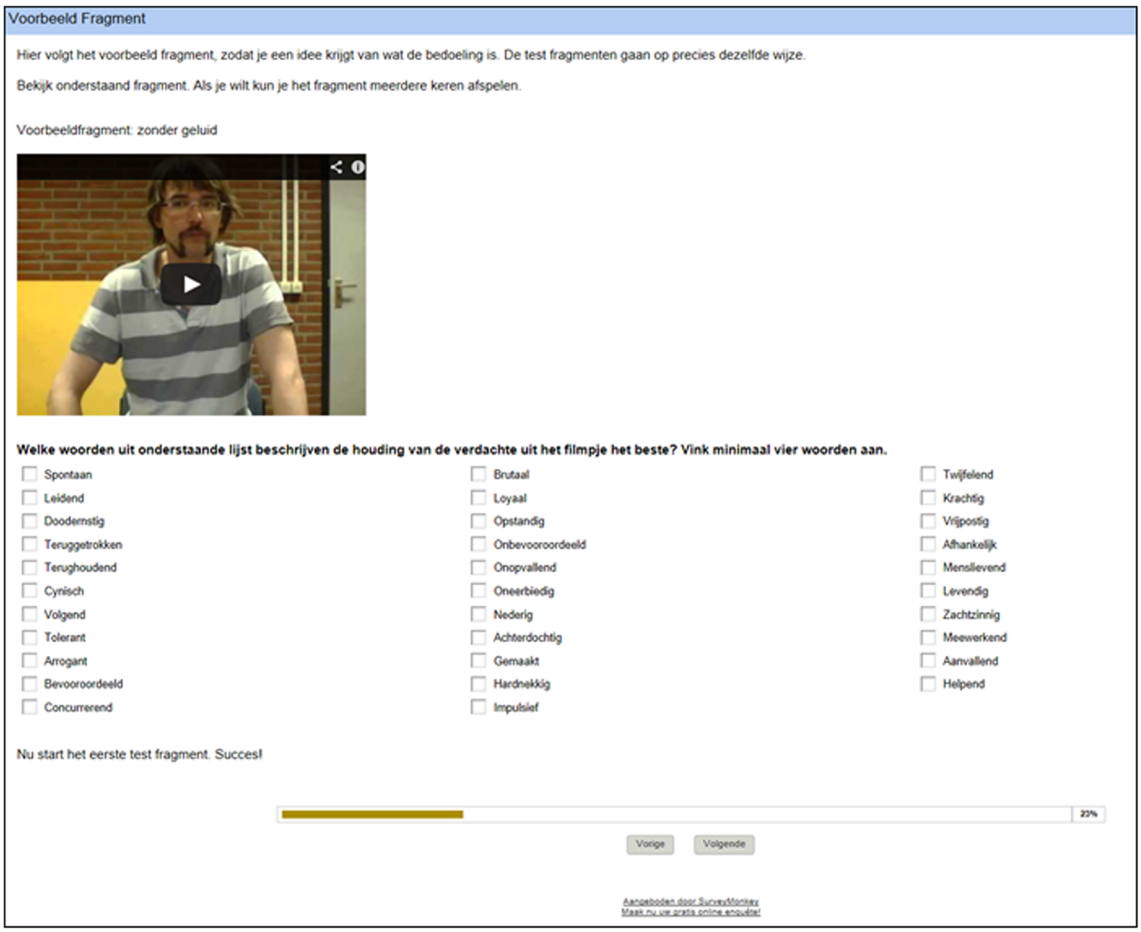

Fig. 4 Screenshot of the survey used to detect how well people recognized the acted stances

5 Results: recognizing stance

Are the acted stances valid, meaning that judges recognize the acted stance? First, we will focus on the distributions of the responses to get a first indication of how well people performed at annotating the videos. Second, the individual judgements will be investigated to see how well individuals recognized the stances. We computed inter-rater agreement. The best recognized videos for each stance have been annotated to extract key poses, gestures and facial expressions that can be used when a conversational agent has to convey a certain stance. Note that when validity is high inter-rater agreement is also high, but a high inter-rater agreement does not imply that the agreed stance coincides with the intended stance; validity can be low for some or all stances.

5.1 Distribution of responses

To get a first indication of how well acted stances were recognized, it was tested whether adjectives belonging to the depicted stance were chosen more often than adjectives from other stances. To adjust for respondents choosing many adjectives when annotating a fragment and therefore having a bigger influence, calculations have been made for each annotation reporting the percentage of adjectives that belong to the different stance categories. The distributions of these percentages will be used in this section.

For each of the four stances that were depicted by the actors, a pie chart was made that shows the mean percentages of annotated adjectives belonging to each stance-category. These pie charts can be found in Fig. 5. The figure gives a first indication of how good respondents were at recognizing the depicted stances. It is striking that stance category SH seems to have been chosen the most by the respondents independent of what stance the actor depicted. The next step was to test whether the differences between chosen stance categories that seem apparent in the pie chart are significant.

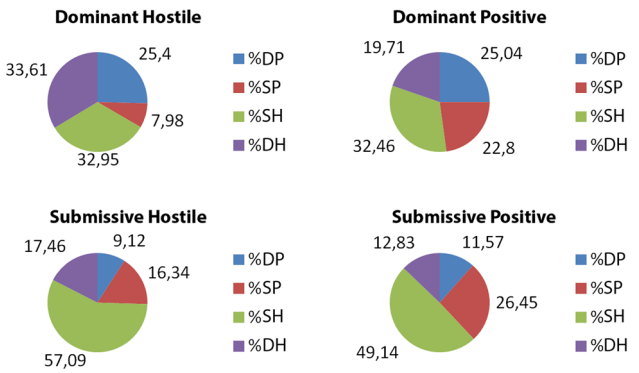

Fig. 5 For each of the acted stances, the pie chart shows the mean percentages of chosen adjectives belonging to the four stance categories

We preformed the Kolmogorov–Smirnov test to test the distribution of the data and concluded that for each acted stance category the data was not normally distributed (all $p < .001$). Therefore, the Kruskal–Wallis test was used to test whether the chosen stance categories differ. Mean scores, standard deviations and test values are shown in Table 3. It can be seen here that for all the acted stance categories there were differences in the percentage annotated adjectives between the chosen stance categories (all $p < .001$). Next, we investigated where these differences were.

To find, for each acted stance category, where the differences between chosen stance categories were, multiple Wilcoxon rank-sum tests were done. It was expected that the stance category that was represented the most would accord with the acted stance category. Therefore it was only tested if this category differed from the other three stance categories that could be chosen and in what direction the difference lay. To counter the inflation of the type-1 error, for each of three comparisons a significance level of .02 was used. The test values of each comparison are shown in Table 4.

It can be concluded that when the stance DP was depicted, the percentage of adjectives chosen that belong to the stance category DP did not significantly differ from the categories DH and SP, but the stance category SH had a higher percentage of chosen adjectives than DP. When the stance DH was depicted, the percentage of adjectives chosen that belong to category DH was significantly larger than the percentages of categories DP and SP. However, it did not differ from SH. When the stance SP was depicted the percentage of adjectives chosen that belong to category SP was significantly larger than the percentages of categories DP and SP, but significantly smaller than category SH. Finally, when the stance SH was depicted, the percentage of adjectives chosen that belong to category SH was significantly larger than all the other categories.

In short, it was expected that the stance that was depicted should also have had the highest percentage of chosen adjectives. This was only the case for stance category SH. Actually, SH had the highest percentage (or shared the highest percentage of chosen adjectives) independently of the acted stance. If SH is not taken into account our expectations are met for the stances SP and DH and partially met for DP (which has a percentage of chosen adjectives that is equal to that of SP and DH). In the next section we will investigate this over representation of SH adjectives.

5.2 Distribution of adjectives

Which adjectives did subjects select for the different categories of fragments? Table 5 shows how many times subjects used each of the 32 adjectives for fragments in the four categories. What does the table show? Compare SH adjectives $ID = 17$ "defiant" and $ID = 24$ "artificial". Both were used

Table 3 For all the stances that respondents could choose from, the mean probabilities, standard deviations and number of observations given a certain acted stance are shown

Acted st.	Chosen stance												Testvalues	
	DP			DH			SP			SH				
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	χ^2	p
DP	0.250	0.244	162	0.197	0.216	162	0.228	0.236	162	0.325	0.254	162	21.4	<0.001
SP	0.116	0.179	157	0.128	0.173	157	0.265	0.246	157	0.491	0.247	157	193.8	<0.001
SH	0.091	0.160	175	0.175	0.207	175	0.163	0.224	175	0.571	0.258	175	274.8	<0.001
DH	0.255	0.224	178	0.336	0.211	178	0.080	0.148	178	0.329	0.243	178	163.4	<0.001

The table also shows χ^2 and p values that test if these mean probabilities differ. The probabilities represent the percentages of annotated adjectives belonging to a certain stance

Table 4 Testvalues are shown for several Wilcoxon Rank Sum tests in order to see which stances differ in percentages annotated adjectives given a certain acted stance

Act. st.	Cho. st.	Chosen stance							
		DP		DH		SP		SH	
		Z	p	Z	p	Z	p	Z	p
DP	DP	x	x	−1.834	0.067	−0.762	0.446	−2.617	0.009
DH	DH	−4.136	<0.001	x	x	−11.424	<0.001	−0.975	0.330
SP	SP	−5.742	<0.001	−5.148	<0.001	x	x	−7.363	<0.001
SH	SH	−14.369	<0.001	−12.394	<0.001	−12.375	<0.001	x	x

frequently. But where artificial was used for all four stances in about the same number of judgements, defiant was used far more for fragments that expressed a DH stance. We also saw that "helping" and "loyal" fitted the DP stance better than "leading", "powerful", and "stubborn", which were used more often to describe DH fragments. Adjectives "gentle" and "humble" fitted the SP stance best. Adjective "reserved" fitted the SH stance best. Finally, adjectives "offensive", "impulsive" and "cheeky" fitted the DH stance best.

Table 6 shows for each of the fragment categories the order of adjectives used in the judgements. The left-most column shows that over all fragments the adjectives "doubting", "reserved", "defiant", "artificial" and "arrogant" were most frequently used. Remarkably all of them belong to the SH segment in Leary's Circumplex (adjective IDs 17–24).

If we look at the SH column of Table 6 we see that all but one (namely "depending") of the 8 SH adjectives that subjects could choose are in the top 8 of the list of most frequently selected adjectives for judgements of SH fragments. The DH column shows that out of the top 5 only 2 adjectives "offensive" and "cheeky" belong to the DH category. In the DP column none of the top 5 adjectives belong to the DP category. The highest DP adjective is "spontaneous" at rank 7. The 4 adjectives ranked 7–10 all belong to this category, but 5 SH adjectives were more frequently used than these 4 DP adjectives. This is remarkable because these two categories are related to two segments of Leary's Circumplex that are diagonally opposite to each other. The SP column has one adjective in the top 5 that belongs to the SP category of adjectives (13. "humble"). Also here 3 of the top 5 adjectives are SH adjectives.

Why did subjects find that the SH adjectives most appropriately described the stance taken by the actor? Is it because of the fact that they were "artificial" behaviours was more apparent than the stance that was intended to be acted out? And was "doubting" the stance that an actor expresses when he/she intends to act a certain stance but doubting how to express this? This raises the question of whether using acted stance fragments is a good idea for studying whether people agree in describing the stance that someone takes. On the other hand it could also indicate that people were biased towards interpreting stances as SH, and that the accompanying adjectives were their default opinion.

5.3 Individual judgements

In total 84 subjects each judged 8 fragments by selecting at least 4 adjectives from a list of 32 that they found most appropriately described the stance acted by the actor shown in the fragment. So in total we have $84 \times 8 = 672$ judgements. Subjects were asked to choose at least 4 adjectives, no maximum was set. Per fragment subjects used 4.6 adjectives in the mean with a maximum of 10 adjectives. Four adjectives were selected 434 times, 5 adjectives 134 times, 6 adjectives 62 times, 7 adjectives 24 times, 8 adjectives 10 times, 9 adjectives 6 times, and 10 adjectives were selected only 2

Table 5 The adjectives (translated to English) ordered per stance and the counts of how many times subjects assigned each adjective to the fragments in each of the four categories

Stance	ID	Adjective	ALL	DP	SP	SH	DH
DP	1	Leading	78	25	10	8	35
	2	Powerful	113	24	7	11	71
	3	Stubborn	106	20	11	26	49
	4	Helping	47	28	9	5	5
	5	Spontaneous	68	28	12	5	23
	6	Lively	57	19	9	4	25
	7	Loyal	51	26	13	6	6
	8	Humane	41	15	15	7	4
SP	9	Unprejudiced	34	11	8	11	4
	10	Tolerant	30	14	4	9	3
	11	Gentle	59	13	32	12	2
	12	Cooperative	108	52	28	15	13
	13	Humble	102	22	51	24	5
	14	Discrete	79	15	27	32	5
	15	Dead serious	83	19	23	13	28
	16	Dependent	66	22	26	13	5
SH	17	Defiant	198	24	19	54	101
	18	Irreverent	120	21	19	45	35
	19	Depending	73	10	32	28	3
	20	Withdrawn	235	38	78	97	22
	21	Arrogant	141	34	17	46	44
	22	Doubting	236	50	96	71	19
	23	Reserved	141	18	49	63	11
	24	Artificial	166	39	38	50	39
DH	25	Cynical	122	37	17	40	28
	26	Compete	38	7	3	8	20
	27	Bold	52	18	7	8	19
	28	Offensive	102	12	6	4	80
	29	Biased	81	20	10	22	29
	30	Impulsive	75	16	9	6	44
	31	Cheeky	115	20	11	25	59
	32	Suspicious	83	15	28	26	14

times. Since the adjectives belong to one of 4 categories of Leary's Circumplex, it is interesting to see how often there was a match between the category of the adjective chosen and the category of the stance acted out in the fragment. A judgement of a fragment by a subject is called:

- Perfect, when *all* the adjectives that a subject has chosen as describing that fragment belong to the same class as the stance that was intended by the actor in the fragment.
- Correct, when there is a unique category with a maximum number of adjectives selected (a unique majority category) and this category is the same as the intended stance of the fragment.
- Semi-correct, when the category that has the maximum number of adjectives chosen is the same as the intended stance of the fragment.
- Wrong, if it is not semi-correct.

Note that when a judgement is perfect it is also correct and when it is correct it is also semi-correct. Thus, a judgement is either semi-correct or wrong and their sum is the total number of fragments of a class. A judgement that has for example 4 adjectives 2 of which are of the intended stance and 2 are of another stance category is semi-correct. It is not correct since it has no unique majority category.

5.3.1 Results for all fragments

Table 7 shows for each of the categories how many times the judgements were perfect, correct, semi-correct, or wrong. The total number of judgements is 672. There are small differences in the numbers of fragments in each of the four categories. From the total of 672 judgements 162 judgements concern DP fragments, 178 concern DH fragments, 157 concern DP, and 175 SH fragments. Table 8 shows how many adjectives subjects assigned to the subsets of fragments of the four different stance categories. For example in total 454 adjectives of the category SH were assigned to the fragments of category SH, but 234 of these SH adjectives were assigned to the fragments of stance category DP. It is clear that by far most of the adjectives selected by the subjects belong to the category SH. This explains the outstanding number of perfect judgements made for the SH fragments (see Table 7).

Table 9 shows the confusion matrix. It shows for each of the stances (rows) how often fragments of that stance were assigned the four classes if we take the stance category with the maximum number of adjectives as the stance assigned. In cases where there was no unique stance category with a majority then the decision is *X* (undecided). From the numbers in Table 9 we computed the precision, recall and F-values (Table 10). The SH and DH (both hostility) categories have clearly higher F-measures than the two categories DP and SP (both positivity). The highest precision was obtained for class DH.

5.3.2 Clustering judgements

To further investigate the structure of the ratings we performed a k-means clustering (with $k = 4$ as there were four stances) of the ratings for every adjective (32 adjectives, so a 32-dimensional binary vector). Every rating was a case (so there were 672 cases) and each fell in one of the four clusters (CL1–CL4). For every case we know the intended stance of the actor, so we can see how often each acted stance occurred in each cluster (Table 11). We gave the cluster the name of the stance that occurred most in that cluster. We compared

Table 6 The adjectives (translated to English) ordered according to frequency of use (descending order) in the 672 judgements for all of the fragments and for the fragments from each of the four categories

ID	Adj(ALL)	ID	Adj(DP)	ID	Adj(SP)	ID	Adj(SH)	ID	Adj(DH)
22	Doubting	12	Cooperative	22	Doubting	20	Reserved	17	Defiant
20	Reserved	22	Doubting	20	Reserved	22	Doubting	28	Offensive
17	Defiant	24	Artificial	13	Humble	23	Withdrawn	2	Powerful
24	Artificial	20	Reserved	23	Withdrawn	17	Defiant	31	Cheeky
21	Arrogant	25	Cynical	24	Artificial	24	Artificial	3	Stubborn
23	Withdrawn	21	Arrogant	11	Gentle	21	Arrogant	21	Arrogant
25	Cynical	5	Spontaneous	19	Depending	18	Irreverent	30	Impulsive
18	Irreverent	4	Helping	12	Cooperative	25	Cynical	24	Artificial
31	Cheeky	7	Loyal	32	Suspicious	14	Discrete	18	Irreverent
2	Powerful	1	Leading	14	Discrete	19	Depending	1	Leading
12	Cooperative	17	Defiant	16	Dependent	32	Suspicious	29	Biased
3	Stubborn	2	Powerful	15	Dead serious	3	Stubborn	25	Cynical
13	Humble	13	Humble	17	Defiant	31	Cheeky	15	Dead serious
28	Offensive	16	Dependent	18	Irreverent	13	Humble	6	Lively
15	Dead serious	18	Irreverent	25	Cynical	29	Biased	5	Spontaneous
32	Suspicious	31	Cheeky	21	Arrogant	12	Cooperative	20	Reserved
29	Biased	3	Stubborn	8	Humane	16	Dependent	26	Competitive
14	Discrete	29	Biased	7	Loyal	15	Dead serious	22	Doubting
1	Leading	15	Dead serious	5	Spontaneous	11	Gentle	27	Bold
30	Impulsive	6	Lively	31	Cheeky	9	Unprejudiced	32	Suspicious
19	Depending	23	Withdrawn	3	Stubborn	2	Powerful	12	Cooperative
5	Spontaneous	27	Bold	1	Leading	10	Tolerant	23	Withdrawn
16	Dependent	30	Impulsive	29	Biased	1	Leading	7	Loyal
11	Gentle	32	Suspicious	4	Helping	27	Bold	14	Discrete
6	Lively	14	Discrete	6	Lively	26	Competitive	13	Humble
27	Bold	8	Humane	30	Impulsive	8	Humane	16	Dependent
7	Loyal	10	Tolerant	9	Unprejudiced	7	Loyal	4	Helping
4	Helping	11	Gentle	2	Powerful	30	Impulsive	9	Unprejudiced
8	Humane	28	Offensive	27	Bold	5	Spontaneous	8	Humane
26	Competitive	9	Unprejudiced	28	Offensive	4	Helping	19	Depending
9	Unprejudiced	19	Depending	10	Tolerant	6	Lively	10	Tolerant
10	Tolerant	26	Competitive	26	Competitive	28	Offensive	11	Gentle

Note that the numbers refer to the *identifing number* of the adjective, not to the count of occurences: adjectives for DP are identified with nrs 1–8, SP 9–16, SH 17–24, and DH 25–32

the F-values of the clustering (Table 12) and the F-values of the assignments by majority vote (Table 10): this showed us (unsurprisingly) that by clustering the F-values went up. This means that the ratings of the judges seem to be structured differently by the clustering compared to the stance categories. With this method we can only see to what extend the judges chose the same adjectives, we cannot investigate whether the different judges chose adjectives from the same stance category. Judges might have disagreed in the adjectives that they chose, but those adjectives might have (largely) belonged to the same stance category.

To test this, we compared the clustering of the adjectives (32-dimensions) to a clustering of a 4-dimensional vector that represented the stance grouping of the adjectives. For every case, we counted how many adjectives from the group of adjectives of the stance were chosen. This resulted in a 4D vector (each stance) with values that ranged from 0 (no adjective from that stance was chosen) to 8 (all adjectives from that stance were chosen). We performed a k-means ($k = 4$) on this 4-dimensional representation of the data. Again, for every case we know the intended stance of the actor and can see how often each acted stance occurred in each cluster (Table 13). We gave the cluster the name of the stance that occurred most in that cluster. We compared the F-values of the two clusterings (Tables 12, 14): the F-values for the 4D clustering were lower for stances DP, SP and DH, and higher

Table 7 The number of times subjects assigned the "correct" stance to the fragments in each of the four categories

CAT	Judgements				Total
	PERF	CORR	SEMICOR	Wrong	
DP	3	28	53	109	162
SP	1	27	47	110	157
SH	22	113	138	37	175
DH	0	52	84	94	178

For explanation of what "correct" means see the main text

Table 8 The number of times subjects assigned the stance adjectives indicating the different stance categories to the fragments in each of the four categories

CAT	Chosen stance			
	DP-A	SP-A	SH-A	DH-A
DP	185	168	234	145
SP	86	199	348	91
SH	72	129	454	139
DH	218	65	274	293

Table 9 The number of times subjects assigned stances to the fragments in each of the four categories or if the chosen stance was undecided

CAT	Chosen stance				
	DP-C	SP-C	SH-C	DH-C	X-C
DP	28	31	43	20	40
SP	13	27	77	7	33
SH	6	15	113	14	27
DH	36	6	49	52	35

Table 10 The accuracy, precision, recall and F-values for each of the four stance categories

CAT	Acc	Precision	Recall	F
DP	0.72	0.34	0.17	0.24
SP	0.73	0.34	0.17	0.24
SH	0.66	0.40	0.65	0.52
DH	0.75	0.56	0.29	0.38

These figures are based on the figures in the confusion Table 9

for SH. So by clustering over the 4D stances the performance dropped. This means that the choices of the judges did not fall into the same stance categories, but were distributed over different stance categories.

5.3.3 Results for sound and mute fragments

There were 336 S-fragment (sound/with audio) judgements, the same as the number of M-fragment (mute/no audio) judgements. Tables 15 and 17 show the scores and the assignment of adjectives for the part of the corpus with sound. Tables 16 and 18 show the scores and the assignment of adjectives for the part of the corpus without sound.

In Fig. 6 the judgements with sound, muted and total are visualised in a bar graph. The total value has been divided by 2 which represents the judgements if sound and mute were to be fully equally distributed. From this graph we can compare the judgements of the S- and the M-fragments and we see that there are no significant differences between the

Table 11 Counts how many acted stances occur in the clusters

CAT	Counts of ratings in each cluster				Total
	CL3-DP	CL4-SP	CL2-SH	CL1-DH	
A-DP	64	30	38	30	162
A-SP	34	88	23	12	157
A-SH	20	81	59	15	175
A-DH	10	10	43	115	178
Total	128	209	163	172	672

The majority of occurrences is the name for that cluster

Table 12 The precision, recall and F-values of the 4-means clustering based on Table 11

CAT	Precision	Recall	F
DP	0.50	0.40	0.44
SP	0.42	0.56	0.48
SH	0.36	0.34	0.35
DH	0.67	0.65	0.66

Table 13 Counts how many acted stances occur in the clusters on the 4(stance)-dimensional data

CAT	Counts of ratings in each cluster				Total
	CL3-DP	CL4-SP	CL2-SH	CL1-DH	
A-DP	45	42	27	48	162
A-SP	18	53	53	33	157
A-SH	9	38	85	43	175
A-DH	44	9	32	93	178
Total	116	142	197	217	672

The majority of occurrences is the name for that cluster

Table 14 The precision, recall and F-values of the clustering based the 4(stance)-dimension data, based on Table 13

CAT	Precision	Recall	F
DP	0.39	0.28	0.32
SP	0.37	0.34	0.35
SH	0.43	0.49	0.46
DH	0.43	0.52	0.47

Table 15 The number of times subjects assigned the "correct" stance to the S-fragments in each of the four categories

CAT	Judgements sound fragm.				Total
	PERF	CORR	SEMICOR	Wrong	
DP	1	15	33	57	90
SP	0	13	23	51	74
SH	11	60	68	17	85
DH	0	25	43	44	87

For explanation of what "correct" means see the main text

Table 16 The number of times subjects assigned the "correct" stance to the M-fragments in each of the four categories

CAT	Judgements Mute Fragm.				Total
	PERF	CORR	SEMICOR	Wrong	
DP	2	13	20	52	72
SP	1	14	24	59	83
SH	11	53	70	20	90
DH	0	27	41	50	91

For explanation of what "correct" means see the main text

Table 17 The number of times subjects assigned stances to the S-fragments in each of the four categories

CAT	Chosen stance sound fragm.				
	DP-C	SP-C	SH-C	DH-C	X-C
DP	15	16	19	10	30
SP	4	13	38	3	16
SH	2	6	60	7	10
DH	13	1	28	25	20

Table 18 The number of times subjects assigned stances to the M-fragments in each of the four categories

CAT	Chosen stance mute fragm.				
	DP-C	SP-C	SH-C	DH-C	X-C
DP	13	15	24	10	10
SP	9	14	39	4	17
SH	4	9	53	7	17
DH	23	5	21	27	15

fragments with and without audio. The percentages of wrong judgements is the same for the fragments with audio and without audio. This holds for all 4 stances.

5.3.4 Results for the theory actors

We had two different conditions in which actors were asked to perform the four stances. Four of the eight actors were recorded in the 'Theory condition'. The other four in the scenario or 'Role play condition'.

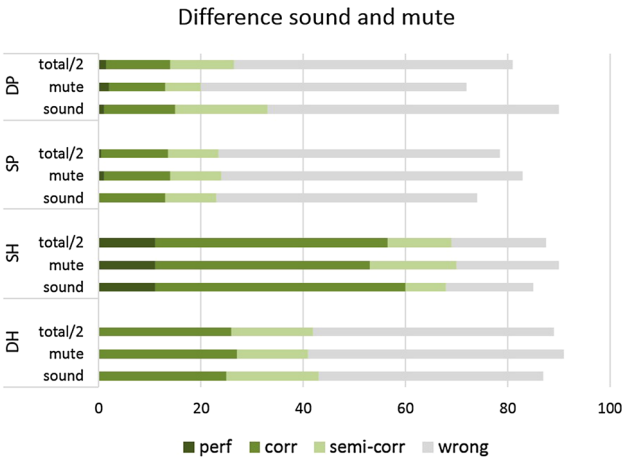

Fig. 6 Difference between mute and sound

Table 19 The counts how many times subjects assigned the "correct" stance to the T-fragments in each of the four categories

CAT	Judgements theory fragm.				Total
	PERF	CORR	SEMICOR	Wrong	
DP	2	20	32	43	75
SP	0	11	19	60	79
SH	19	74	83	10	93
DH	0	31	43	46	89

For explanation of what "correct" means see the main text

Are there differences between these two groups of actors if we look at how many judgements were correct? Or, in other words did subjects recognize the intended stances better when this stance was acted in the Theory condition than when the stance was acted in the Role play condition? Half of the judgements (336) involve actors in the T-condition, the other half in the R-condition.

The Tables 19 and 21 show the results for the T-actors. The Tables 20 and 22 show the results for the R-actors. In Fig. 6 the judgements with sound, muted, and total are visualised in a bar graph. The total value is divided by 2 which represents the judgements if sound and mute were to be fully equally distributed. If we compare the figures in Tables 19 and 20 we see that of the 22 perfect judgements involving an acted SH stance, 19 were performed in the Theory condition. Only 3 in the Role play condition. In Fig. 7 the judgements for the Theory condition, Role play condition and total are visualised in a bar graph. The total value is divided by 2 which represents the judgements if Theory and Role play were to be fully equally distributed. As can be seen here it appears that overall the Theory conditions induces acted stances that are better recognized than those in the Role play condition.

Table 20 The counts how many times subjects assigned the "correct" stance to the R-fragments in each of the four categories

CAT	Judgements role play fragm.				Total
	PERF	CORR	SEMICOR	Wrong	
DP	1	8	21	66	87
SP	1	16	28	50	78
SH	3	39	55	27	82
DH	0	21	41	48	89

For explanation of what "correct" means see the main text

Table 21 The counts how many times subjects assigned stances to the T-fragments in each of the four categories

CAT	Chosen stance theory fragm.				
	DP-C	SP-C	SH-C	DH-C	X-C
DP	20	9	17	8	21
SP	10	11	39	2	17
SH	1	3	74	6	9
DH	9	1	35	31	13

Table 22 The counts how many times subjects assigned stances to the R-fragments in each of the four categories

CAT	Chosen stance role play fragm.				
	DP-C	SP-C	SH-C	DH-C	X-C
DP	8	22	26	12	19
SP	3	16	38	5	16
SH	5	12	39	8	18
DH	27	5	14	21	22

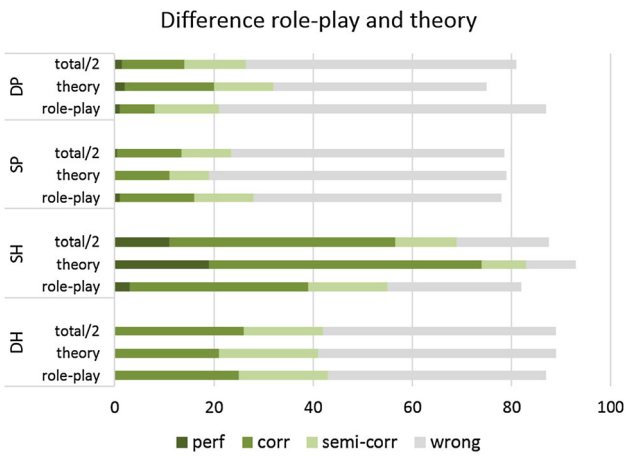

Fig. 7 Difference between theory and role-play groups

5.3.5 Preliminary conclusion individual judgements

The results presented in this section show that there is a clear correlation between the stance of the fragments and the adjectives chosen by the total of all subjects in their judgements of these fragments. This is clear from the distribution, Table 5. The observed frequencies on the main diagonals of these tables are always considerably larger than their expected values. The $\chi^2(df = 9)$ values are respectively 383, 254 and 171 all with $p \ll 0.0001$. Of course this is as it should be. Ideally all values should be on the main diagonal. On the other hand there were many judgements in which subjects chose adjectives that belong to a different category than the stance category that was intended by the actor. Up to now we have only analysed which adjectives the group of all subjects used to describe the fragments of the various stance classes. It is quite possible that in cases where subjects did not recognize the stance as it was intended, they were at least in agreement with each other.

5.4 Inter-annotator agreement

Did subjects agree on the use of adjectives for the different fragments? If subjects did not agree in their accounts of the stance taken by the actors in the fragments they judged then it is difficult to say what the stance was that the actor took. The meaning of "content" in the discipline of content analysis is not always clear [26]. Here content is something of the interaction between what is presented, in this case the behaviour shown in the video fragments and the subject who had the task to describe the stance taken by the actor. There is no "golden truth". If most of the subjects who judged a certain fragment judged this as a *submissive* stance then this is something we have to take as "content" even if the stance that was intended by the actor was more of the type *opposing* than *submissive*. We analysed the judgements for inter-rater agreement. Our "coding task" had the following properties.

- There was a large number of annotators (84).
- Not all annotators annotated all fragments. There was a total of 64 fragments: 8 actors, acted each 4 stances, and we have all recordings with and without audio, makes $2 \times 8 \times 4$ fragments. Each of the annotators labelled 8 fragments. The fragments were assigned to the annotators in a random way, but ensuring that not all fragments belonged to the same category.
- The label set used in the annotation task is large. A subject could label a fragment with any subset of the set of 32 adjectives, with the restriction that it contained at least four adjectives.

Because of these properties we used Krippendorff's α agreement method for computing a reliability measure ([26], p. 222). For a thorough discussion about the various methods and measures for inter-rater agreement see [2]. We applied the method for many observers, many nominal categories, and for missing values ([26], p. 232).

Since the annotators assigned a set of adjectives to each fragment, category labels in this annotation task consist of sets of adjectives. However, since the number of potential labels was quite large (potentially as many as there were subsets—with at least 4 elements—of a set of 32 elements) and many of them had not been used, we decided not to use sets of adjectives, but sets of stance categories. Note that annotators did not know the stance categories of the adjectives; the adjectives were given in a random order. Theoretically we now had $2^4 - 1 = 15$ different labels. They correspond to the non-empty subsets of the four stance categories.

Krippendorff's α allows us to plug in a distance metric on the label set. We used a difference metrics based on the similarity measure on sets known as *Dice coefficient*, see [2,16,30]. Suppose two annotators assigned sets C_1 and C_2 each containing 4 adjectives to a certain fragment. In each of the two sets 2 adjectives belonged to stance category X, the other two belong to Y. This means that both annotators were in a sense uncertain about the stance expressed. But the metric does not penalize this as disagreement. To give another example: the distance between the sets $\{DP, DH, SP\}$ and $\{DH, SP\}$ is 0.2. The results of the inter-rater agreement analysis are shown in Table 23. It shows the α values for the whole class of fragments and for the class of S-fragments (with audio) and the class of M-fragments with muted audio. These values are low. There was slightly more agreement on the fragments with audio than there was on the fragments without audio. Clearly, the Theory play judgements had a higher inter-rater agreement. We also computed α for parts of the corpus containing only fragments of a certain intended stance, see Table 24. This table also shows the α values for the corpus without the parts containing fragments of a specific stance. The exceptional values for the *DT* fragments are remarkable. Remember that this was the class that also had the highest precision value. DT stance behaviour is easier to recognize (and perform!) than the other types of stances.

Table 23 The α values computed for fragments with and without sound and for the role play and theory fragments using Krippendorff's method with Dice metrics for distances between values

ALL	α-Audio		α-Condition	
	Mute	Sound	Role play	Theory
0.22	0.21	0.23	0.15	0.27

Table 24 The α values computed for the fragments of each acted stance and for all fragments excluding the fragments of a specific acted stance using Krippendorff's method with Dice metrics for distances between values

α-Stance categories							
DP	−DP	SP	−SP	SH	−SH	DH	−DH
0.12	0.23	0.08	0.24	0.03	0.21	0.22	0.15

5.5 Were some actors better than others?

We saw that stances acted in the Theory condition were better recognized and had a higher inter-rater agreement than the stances acted in the Role play condition. In each condition 4 actors performed the stances. Now we will look at individual actors. Were some actors better than others in the sense that the stances they performed were easier to recognize by the subjects? To answer this question we computed a score for each of the actors. For each of the actors we looked at the judgements in which the actor acted. If the judgement was perfect we added 3 points to the score, if it was correct we added 2 points to the score, if it was semi-correct we added 1 point to the score. Since all actors were involved in the same number (84) of judgements we did not have to normalize these scores. The resulting scores are in Table 25. The T-actors are the actors that acted in the Theory condition, the R-actors are those that acted in the Role play condition. Actor T04 scored significantly higher than the mean score and actor R04 scored lower than the mean. What is the impact of these two actors on the *alpha* values? If we remove all judgements with R04 α slightly raises from 0.22 to 0.25. If we remove T04 α becomes 0.19. If we only take the fragments with actor T04 α raises to 0.38. Our analysis confirms that some actors were better than others and that good acting had a significant impact on inter-annotator agreement.

Table 25 The scores and α reliability values for each of the 8 actors

Actors in theory-condition							
T01		T02		T03		T04	
Score	α	Score	α	Score	α	Score	α
72	0.16	90	0.22	58	0.03	114	0.38
Actors in role play-condition							
R01		R02		R03		R04	
Score	α	Score	α	Score	α	Score	α
61	0.21	68	0.24	62	0.04	43	0.01

Since fragments were assigned randomly to subjects there is a chance that the positive and negative scores for T04 and R04 were due to the subjects, not to the actors. In order to cancel this out we looked at those judgements by subjects that both annotated the same stance by the same actors (for actors T04 and R04). Do these judgements differ in quality if we vary the subject or if we vary the actor?

The data contains 8 subjects that annotated both actors R04 and T04 acting stance SH, 3 subjects that annotated both actors acting stance SP, 7 for stance DT and 6 for stance DP. In total 24 different subjects annotated the two actors acting the same stance. For each of these 48 judgements we computed the scores and we analysed the results. The scores for R04 has a mean of 0.42 (SD 0.88) and for T04: mean is 1.42 (SD 0.93). A paired t test comparing scores for R04 and T04 on the 24 pairs of judgements of the same stances by the same subjects gives: $t(23) = 4.796$ ($p \ll 0.0001$). In all but one case the judgement of a subject has a higher score with actor R04 than with actor T04. In all other cases T04 scores equal (9 times) or higher (14) than R04. This gives sufficient evidence to rule out that the higher scores for T04 compared to those for R04 are due to the judges assigned to them. Table 25 also contains the α reliability values for the 8 parts of the corpus divided per actor. Figure 8 shows the relation between scores and α values. It shows that the ratio between scores and α values varies considerably. For actors $R01$ and $R02$ they are much higher than for R03. The Spearman correlation between α and score equals 0.833 (significance $p < 0.05$, 2-tailed). Overall, there is a reasonable correlation between the inter-annotator agreements and the validity. But as the correlation graph shows, for some actors (e.g. R03) a higher score (validity: agreement between judgement and intended stance) goes with a low inter-annotator agreement and for others (e.g. R01, R02) a lower score (validity) goes with a higher inter-annotator agreement meaning relatively more annotators agree on the stance they see but it is not the stance as it was intended. We see that an actor being good has two different senses: he performs the stance that he was asked to act, or he performs a stance that is recognized by a majority of observers. We see that in the mean the T-actors have higher scores as well as higher inter-annotator agreement than the R-actors.

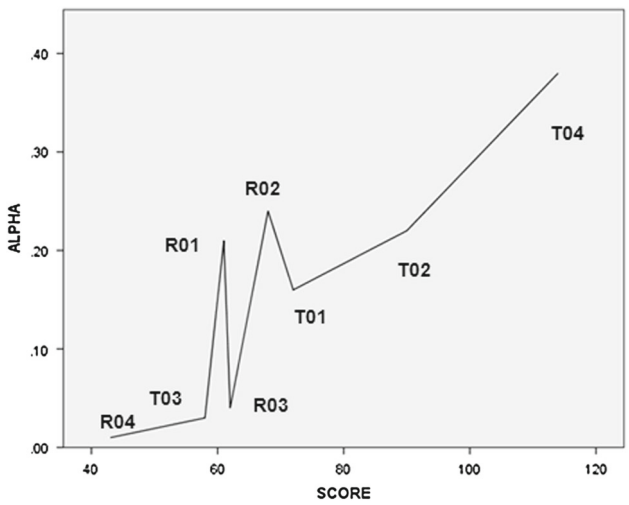

Fig. 8 Judgement scores and inter-annotator agreement for each of the 8 actors

Table 26 Selected fragments

Stance	English description	Actor nr.
DP	Dominant–positive	4
SP	Submissive–positive	5
SH	Submissive–hostile	2
DH	Dominant–hostile	4

5.6 Best fragments

This research was conducted to contribute to a project with a wider perspective. Namely to bring forth a conversational agent that can be used to train interrogation skills for police men and women. The part this research will take in the bigger project is to try to describe certain postures which could be depicted by a conversational agent. If these different postures are valid, they can be used to provoke certain reactions according to Leary's interpersonal stance relations ([27]). As described before, the relation between depicted stance and perceived stance seems very weak. This is why it could be difficult to clearly define a typical and valid posture that depicts a certain stance. Nevertheless we will try to qualitatively describe the best fragment of each depicted stance. In order to determine which fragments are the best, all 4 stances of all 8 actors, which represents all fragments, were judged and plotted. The judgement of these fragments was done by inter-rater agreement and the score system as used before. This plot is shown in Fig. 10. For practical reasons the actors are numbered consecutively where actor number 1 till 4 represent T01 till T04 and 5 till 8 represent R01 till R04. As can be seen in this plot alpha reliability values are very low. This is probably mainly caused by the small number of respondents on each separate fragment. As described by [26] these values are far from relevant and therefore will not be used in the judgement of the fragments. When only taking the judgement scores in account the selected fragments can be seen in Table 26. Figure 9 shows stills for the best actors for the four stances. There is a similarity observable between the stills of different actors for the same stance. For example, raised eye-brows for DP, downward gaze for SP, folded arms for SH, and raised arms for DH.

Fig. 9 The best actors for each of the four stances. Observe the apparent similarity across different actors within the same stances

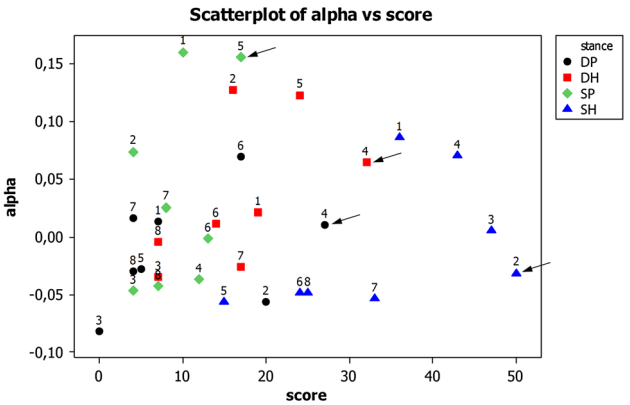

Fig. 10 Judgement scores and inter-rater agreement for each of the 8 fragments

6 Second perception test

Some of the actors in the previous study mentioned they found the task of acting out an interpersonal stance difficult and unnatural. This might have influenced the judgements in such a way that the SH-adjectives were chosen as the most appropriate for most acted stances. Their behaviour felt 'artificial' which was one of the adjectives describing the SH stance. The fact that the behaviour was 'artificial' might have been more apparent than the stance that was intended. The question is whether the (lack of) expertise of our actors might have obfuscated the intended stance with behaviour that is interpreted as SH. To address this issue, we repeated the study using clips taken from TV-shows that feature police interrogations. The thought here was that the professional actors in these clips are better at acting. In addition, we asked raters explicitly about the spontaneity of the behaviour of the actor in the fragment.

6.1 Professional actor fragments

We selected fragments from TV-series. The interpersonal stance of the suspect in each fragment was determined from the content of the entire episode. This could mean anything from the content of what they said to explicit comments made by the characters in the episode. Observing the (non verbal) behaviour of the actors, we kept in mind the typical stance behaviours from the literature, see Table 2. We categorised the stance in these fragments using our best judgement, however we have seen from previous work on stance judgements that this subjective task often has a low inter-rater agreement (e.g. [35]). In other words, we and the participants in our study might not agree on the stance that is portrayed in a fragment. This introduces uncertainty about which stances the fragments that we selected would actually depict according to a majority vote of multiple observers. The thought here was that a majority vote on stance would be closer to the stance that is portrayed than an expert opinion. We assessed which stance was actually depicted in the fragments, see Sect. 6.3.

To ensure that the fragments were similar and comparable to the fragments used in the experiment described earlier, the fragments had to meet three criteria: the suspect is the only one in the fragment, the suspect is being interrogated (seated in a room), and the length of the fragment was similar to the acted stances in the previous study (3–10 s). Figure 11 shows stills of some of the video fragments from TV-series used in the study. The actors show different stances (see caption). We selected four fragments for every stance: three from professional actors and the best recognized fragment from the previous study with our amateur actors. These fragments allow us to compare both data-sets.

Fig. 11 Stills from the fragments with the professional actors. **a** Dominant–hostile expression: tilt head up with a gaze down toward the interrogators; **b** dominant–hostile posture: asymmetrical, space-filling and distant posture; **c** submissive posture: shrinking posture with bent spine; **d** hostile expression: expression of contempt; **e** dominant expression: expressive face with extreme brow raising; **f** submissive expression and posture: expression of sadness and self-touch; **g** friendly expression: smile; **h** hostile expression: cross-eye gaze with eyelid raising

6.2 Method

Participants observed each fragment and rated them on the two interpersonal stance dimensions, dominance and affiliation, and also on spontaneity. Instead of having participants label the fragments with 32 adjectives we opted for rating on the stance dimensions to mitigate concerns about the ambiguity that using adjectives might have introduced, see Sect. 5.1. To assess the quality of acting we asked how spontaneous the behaviour in the fragment appeared. We used 5-point Likert scales. The labels on the scales were: very dominant (5)–very submissive (1) for dominance, very friendly (5)–very hostile (1) for affiliation, and very spontaneous (5)–very acted (1) for spontaneity.

6.3 Fragments' stance

In total 65 participants (aged: *mean* = 25.4, *min* = 14, *max* = 50, and *SD* = 6.2. Gender: 31 female.) judged the 16 fragments on 5-point Likert scales on the three dimensions: dominance, affiliation, and spontaneity. If the mean of the judgements of all participants was above or below the midpoint (which is 3) we classified the fragment in the respective category. For example, if for a fragment the mean on the dominance scale was above 3 it was rated as dominant, whereas the mean was below 3 the fragment was rated as submissive. This analysis can show us the stance that was depicted in each fragment. Table 27 shows our prediction of stance versus the result of this categorization of the fragments based on the judgements of the participants. We concluded that we had 3 fragments that depict a DP stance, 6 DH, 4 SP, and 3 SH (see also Fig. 12). Note that this includes the fragments from our amateur actors, see Sect. 6.4.

Table 27 Our prediction of stance versus the categorization of the fragments based on the judgements of the participants

	Outcome			
	DP	DH	SP	SH
Predicted				
DP	3	0	1	0
DH	0	4	0	0
SP	0	1	3	0
SH	0	1	0	3
Total	3	6	4	3

6.4 Amateur actors

We know the intended stance from the fragments with our amateur actors. Fragments 5, 9, 13, and 17 feature our amateur actors, see Table 28. Fragment 5 was acted with a DP stance and the mean rating of dominance was above 3 meaning it was rated as dominant, and the mean rating of affiliation was above 3 meaning it was rated as positive: a DP stance. Fragment 9 was acted with a DH stance and the mean rating of dominance was above 3 meaning it was rated as dominant, and the mean rating of affiliation was *below* 3 meaning it was rated as hostile: a DH stance. Fragment 13 was acted with a SP stance and the mean rating of dominance was below 3 meaning it was rated as submissive, and the mean rating of

Table 28 Mean ratings for the fragments for dominance, affiliation, and spontaneity

Fragm.	Dom.	Aff.	Stance	Spont.
	Mean	Mean	Result	Mean
2	4.03	3.12	DP	3.31
3	3.20	3.40	DP	3.06
4	2.62	3.14	SP	3.42
5-DP	3.65	3.66	DP	2.65
6	3.97	2.48	DH	3.05
7	3.86	2.46	DH	3.03
8	3.08	2.62	DH	3.22
9-DH	3.55	2.94	DH	2.45
10	2.20	3.05	SP	3.14
11	3.18	2.89	DH	2.97
12	2.42	3.02	SP	3.09
13-SP	2.34	3.42	SP	2.94
14	2.31	2.75	SH	3.06
15	3.69	2.17	DH	3.14
16	2.62	2.46	SH	3.05
17-SH	2.63	2.88	SH	3.00

The fragments with our amateur actors (fragments 5, 9, 13, and 17) have the stance as intended by the actor displayed. Note that fragment 1 was the example fragment and it is not included in the analyses

affiliation was above 3 meaning it was rated as positive: a SP stance. Fragment 17 was acted with a SH stance and the mean rating of dominance was below 3 meaning it was rated as submissive, and the mean rating of affiliation was below 3 meaning it was rated as hostile: a SH stance. The mean ratings match the stance that was intended by the amateur actors for all of these fragments.

We were concerned that the expertise of our actors might have influenced the perceived stance. In the previous perception study we did not explicitly ask our participants to rate the quality of acting. In this experiment we did, see Table 28, and it showed that none of our amateur actors were rated with a mean over 3 points (the mid-point) for *spontaneity*. From the fragments with professional actors only 2 fragments were rated with a mean score below 3 and 10 were rated with a score of 3.03 or higher. This means that the behaviour of the amateur actors was rated as 'acted' and not as spontaneous, where the behaviour in most professional fragments was rated as spontaneous rather than acted. In the previous experiment we found that our amateur actors were described most with adjectives from the SH stance, see Sect. 5.1. The results make us conclude that the expertise of our actors might have influenced the perceived stance. This conclusion is in line with earlier findings from [9] who suggest that professional actors may provide a more natural representation of interpersonal emotions avoiding exaggeration or caricature behaviours.

6.5 Spontaneity and inter-rater confusion

We can further investigate the relation between the expertise of the actor and the judgements on dominance and affiliation. For this we compare the ratings on spontaneity with the standard deviation on dominance and affiliation for all fragments. The correlation between these measures tells us something about the influence the spontaneity had on the clarity of the acted behaviour. The reasoning is that 'unclearly' acted behaviour will lead to a larger deviation as raters have to guess, they disagree more when the behaviour of an actor is inconsistent or conflicting.

Correlation and regression analyses were conducted to examine the relationship between the predictor spontaneity (quality of acting) and deviation of dominance and affiliation (confusion between raters). Figures 13 and 14 show the scatter-plots for spontaneity and the deviation of affiliation and dominance. It seems that indeed there was a trend (albeit weak) that an increase in spontaneity decreased the deviations. Table 29 shows that spontaneity significantly predicted the confusion for dominance ($p < 0.05$). However, spontaneity was not a significant predictor for the deviation of affiliation ($p > 0.1$).

6.6 Clustering of ratings

To further investigate the structure of the ratings we again performed a k-means clustering (with $k = 4$ as there were four stances) on the ratings on dominance, affiliation and spontaneity for each of the fragments. There were 17 clips

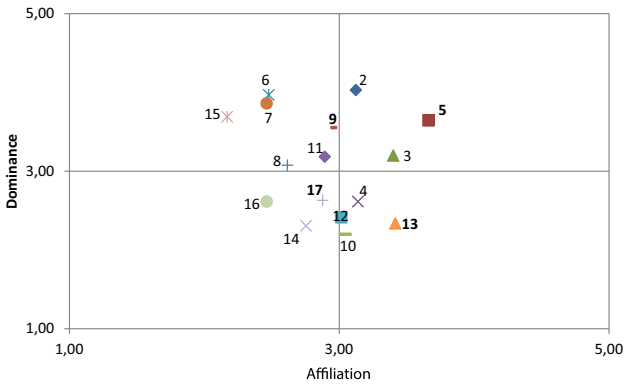

Fig. 12 The location on the interpersonal circumplex of all the fragments based on the mean ratings on dominance (y-axis) and affiliation (x-axis). Note that fragment 1 was used to familiarize participants with the procedure and is excluded from analyses. Fragments 5, 9, 13, and 17 are the amateur actors and are displayed *bold*

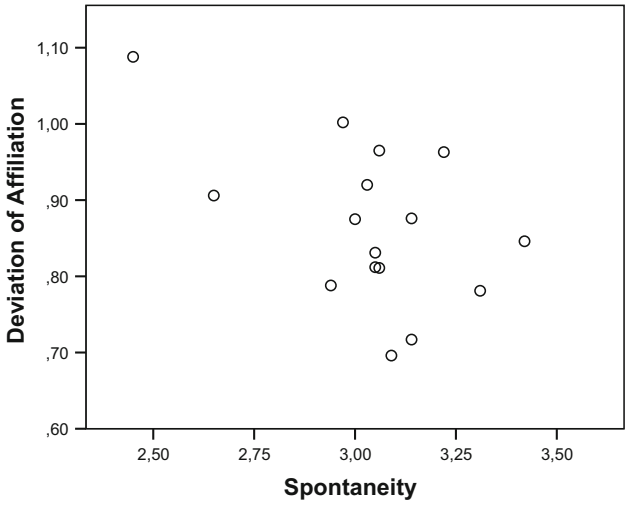

Fig. 13 The scatter-plot for the spontaneity and the deviation of affiliation for all fragments

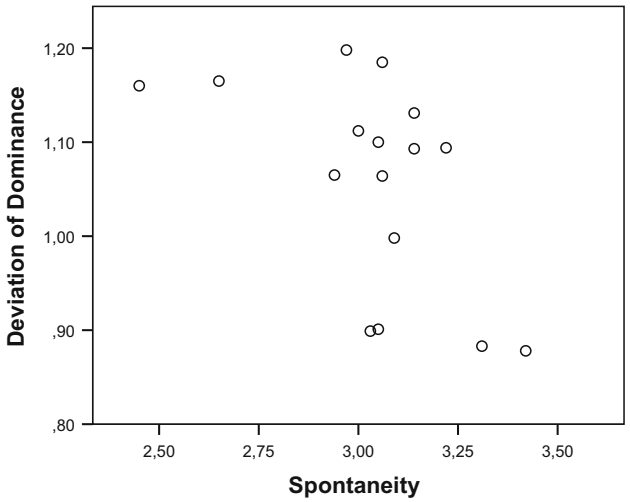

Fig. 14 The scatter-plot for the spontaneity and the deviation of dominance for all fragments

Table 29 Spearman's rho (data is non-parametric) correlations for spontaneity and rater confusion measured by the deviations on affiliation and dominance

	Dev. of aff.	Dev. of dom.
Spont.		
Corr. coef.	−0.398	−0.517*
Sig. (2-tailed)	0.127	0.040
N	16	16

* ($p < .05$)

rated by 65 judges resulting in 1105 cases that were rated on three scales (3-dimensions). Earlier in this section we determined the stance that was *perceived* in each clip. We counted how many times each stance occurred in each cluster. However, we did not have an equal number of clips for each

Table 30 Ratio of counts of judgements in each cluster divided by total judgements for that stance (see main text for details)

CAT	Cluster judgement fractions			
	CL1-DP	CL3-SP	CL2-SH	CL4-DH
DP	0.523	0.036	0.185	0.256
SP	0.265	0.269	0.377	0.088
SH	0.215	0.236	0.385	0.164
DH	0.422	0.075	0.158	0.345
Total	1.426	0.616	1.104	0.854

Table 31 The precision, recall and F-values of the clustering, based on Table 30

CAT	Precision	Recall	F
DP	0.367	0.523	0.431
SP	0.437	0.269	0.333
SH	0.348	0.385	0.366
DH	0.404	0.345	0.372

stance. Therefore we divided the count of how many times the stance occurred in each cluster by how many fragments of that stance were rated, giving the ratio of the counts in the cluster and the occurrences of the stance (Table 30). We gave the cluster the name of the stance that occurred (relatively) most in that cluster. The precision, recall and F-values for this analysis (Table 31) were similar to those for the amateur actor ratings. This means that the confusion of raters was similar for amateur and professional actors.

7 Conclusion

The background idea of the stance perception studies is that there are typical "tiny behaviours" that humans make in a conversation that together make up how their interpersonal stance is perceived by the observer. The idea is quite common, see the references in Table 2. The analyses of our perception studies show once again that there is a complex relation between the isolated observable elements (posture, gesture, or facial expressions) on the one hand and the perceived stance on the other.

The results of the annotations show a clear correlation between the stances that were acted in the videos and the adjectives that were chosen in the judgements. However, there were many judgements in which subjects chose adjectives that belonged to another category than the stance category that was intended by the actor. We see that inter-annotator agreement is low ($\alpha = 0.22$). Other studies that looked at inter-annotator agreement in a stance annotation

task using Leary's Circumplex have already shown that this is a difficult task (see, e.g., [35,45]).

The stance that was seen most in the first perception study is submissive–hostile (SH) independent of the stance that was intended by the actor. This can first of all be because the interpersonal stances in the videos were acted and several actors commented that the task felt unnatural to them which could have influenced the naturalness of their acting. Secondly, this could indicate that raters have a default opinion about the clips they are judging. The raters were explained they were going to watch suspects in a police interrogation. Given this context raters might have been biased in their observations thinking that suspects would have acted hostile and submissive towards a police officer, since they were being charged of doing something wrong by a dominant public figure. This research therefore gives a first indication that information about the setting in which communication takes place, can guide people in their observations. If the found bias towards SH really indicated a bias provoked by the setting or if this was a resultant of using actors, needs to be further investigated. It should also be investigated if different biases are found when different communicative settings are used. In this experiment audio did not add necessary information for annotating interpersonal stance. It has to be noted that some videos contained silent acting. The judgements of actors in the theory-condition did seem to differ from the judgements in the role play-condition. For most stances, fragments with actors from the theory-condition seem to be better recognized. This is most obvious with the acted stance SH where 19 of 22 perfect judgements are in the theory-fragments. It could be of influence that the actors in the theory-condition had the exact same list of adjectives in their instructions as the list that was used in the survey. Some actors were better than others in the sense that they put the stance they intended to show on stage better. Others were better in the sense that the stance they acted was recognized by more spectators. The ratio between validity and inter-annotator agreement differed per actor. It is striking to see that most fragments where the acting was exaggerated were recognized best. For making the virtual suspect this is acceptable as the interrogation game tries to familiarize police trainees with the effects of Leary's theory.

We have seen that the expertise of an actor can influence the perception of his acted behaviour. Fragments of professional actors were rated as more spontaneous than fragments of amateur actors who were rated as more 'acting'. Furthermore, we have seen that the quality of acting (spontaneity) influences the agreement between annotators. Fragments that were rated more spontaneous tended to have lower standard deviations on the ratings of dominance and affiliation. This effect was significant for the deviation of dominance. If one makes an argument for using actors when trying to obtain a ground truth of behaviours, we should be careful to see whether this behaviour is also perceived and interpreted by others to be the intended behaviour. For making a virtual suspect it does not necessarily matter if the actor is experienced or 'good', what matters is that the behaviour is recognized by independent observers. Their judgement should be leading in determining what behaviour to use to model the virtual suspect's behaviour. The perception of the behaviour of this virtual character should then be evaluated. The concern that exists for a human actor also exist for the virtual actor: the intended stance might not be the stance that is perceived by independent observers.

When stills from the best recognized fragments are compared, similarities within stance categories and differences between these categories are apparent, see Fig. 9. In summary, it can be seen that dominant postures are upright with a gaze straight at the conversational partner while submissive postures are more closed with a gaze away from the conversational partner.

The most valuable lesson learned from these studies is that it is hard to act a stance and -maybe even more valuable- that observers often see diverse aspects in the behaviour of someone. People apparently often show a mixture of stances and it depends on the perspective taken by the judge as to which aspect of the suspect's behaviour determines how his stance is perceived.

7.1 Future work

The fact that one particular stance (SH) in the judgements of our participants was favoured warrants more research. Perception studies should be performed in different settings, where the setting might lead to stereotypical thinking. It should be investigated whether these settings lead to biases in judgements as was found in this study in the context of a police interrogation. In this line of research the behaviour should be natural using professional actors or preferably even naturally occurring behaviour in order to rule out the explanation that the stance preference was due to 'overacting'.

In [4] it is argued that "it would be worthwhile to systematically investigate the similarities and differences of emotional expressions produced more or less intentionally (in everyday life and/or in the laboratory). This could involve comparing acted portrayals with less 'controlled' expressions, recorded under conditions that would not promote emphasis or suppression of expressions for the benefit of a receiver." A similar investigation is needed for interpersonal stance.

After the Emotion Recognition in the Wild Challenges [15] that aimed at the recognition of emotions from faces in audio-video clips it seems to be time for an Interpersonal Stance Recognition in the Wild Challenge that aims at the classification of stances expressed by suspect and interrogator in video clips that mimic real-life conditions; the conditions in the interrogation room. This would contribute to

the development of support tools for the (real-time) analysis of interrogative interviews as well as to behaviour modelling for the generation of behaviour that expresses in a believable way the interpersonal stance taken by a virtual suspect character or a virtual interviewer in an interview simulation training environment.

Acknowledgments The research reported here is supported by the Dutch COMMIT project P2: "Natural Interaction for Universal Access".

References

1. Argyle M (1988) Bodily communication, 2nd edn. Methuen, London
2. Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. Comput Linguist 34(4):555–596
3. Ballin D, Gillies M, Crabtree B (2004) A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In: 1st European conference on visual media production (CVMP). IEE, London, UK
4. Bänziger T, Mortillaro M, Scherer KR (2012) Introducing the geneva multimodal expression corpus for experimental research on emotion perception. Emotion 12(5):1161–1179
5. Bänziger T, Scherer KR (2007) Using actor portrayals to systematically study multimodal emotion expression: the gemep corpus. In: Affective computing and intelligent interaction. Springer, Berlin, pp 476–487
6. Baranyi P, Csapó A (2012) Definition and synergies of cognitive infocommunications. Acta Polytech Hung 9(1):67–83
7. Beune K, Giebels E, Sanders K (2009) Are you talking to me? Influencing behaviour and culture in police interviews. Psychol Crime Law 15(7):597–617
8. Birdwhistell R (1970) Kinesics and context. University of Pennsylvania Press, Philadelphia
9. Busso C, Narayanan S (2008) Recording audio-visual emotional databases from actors: a closer look. In: Second international workshop on emotion: corpora for research on emotion and affect, international conference on language resources and evaluation (LREC 2008), pp 17–22
10. Cafaro A, Vilhjálmsson HHi, Bickmore T, Heylen D, Jóhannsdóttir KR, Valgardhsson GS (2012) First impressions: users judgments of virtual agents personality and interpersonal attitude in first encounters. In: Intelligent virtual agents. Springer, Berlin, pp 67–80
11. Carney DR, Hall JA, LeBeau LS (2005) Beliefs about the nonverbal expression of social power. J Nonverbal Behav 29(2):105–123
12. Chollet M, Ochs M, Pelachaud C (2012) Interpersonal stance recognition using non-verbal signals on several time windows. In: Proceedings workshop affect, compagnon artificiel, interaction
13. Chollet M, Ochs M, Pelachaud C (may 2014) Mining a multimodal corpus for non-verbal behavior sequences conveying attitudes. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA)
14. Dael N, Mortillaro M, Scherer KR (2012) Emotion expression in body action and posture. Emotion 12(5):1085–1101
15. Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: Proceedings of the 16th international conference on multimodal interaction, ICMI '14, pp 461–466. ACM, New York, NY, USA
16. Dice L (1945) Measure of the amount of ecologic association between species. Ecology 26(3):297–302
17. Douglas-Cowie E, Cowie R, Cox C, Amir N, Heylen D (2008) The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In: Proceedings of the international conference on language resources and evaluation (LREC), pp 17–22
18. Ekman P, Rosenberg EL (1997) What the face reveals: basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, Oxford
19. Endrass B, André E (2014) Integration of cultural factors into the behavioural models of virtual characters. In: Stent A, Bangalore S (eds) Natural language generation in interactive systems. Cambridge University Press, Cambridge, pp 227–251
20. Gifford R (1994) A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. J Person Soc Psychol 66(2):398–412
21. Gillies M, Ballin D (2003) A model of interpersonal attitude and posture generation. In: Intelligent virtual agents. Springer, Berlin, pp 88–92
22. Hess U, Thibault P (2009) Why the same expression may not mean the same when shown on different faces or seen by different people. In: Affective information processing. Springer, Berlin, pp 145–158
23. Inbau FE, Reid JE, Buckley JP, Jayne BC (2013) Criminal interrogation and confessions, 5th edn. Jones & Bartlett Learning, Burlington
24. Kenny PG, Hartholt A, Gratch J, Swartout W, Traum D, Marsella SC, Piepol D (2007) Building interactive virtual humans for training environments. In: Interservice/industry training, simulation and education conference (I/ITSEC)
25. Kleinsmith A, Bianchi-Berthouze N (2013) Affective body expression perception and recognition: a survey. IEEE Trans Affect Comput 4(1):15–33
26. Krippendorff K (2004) Content analysis: an introduction to its methodology, 2nd edn. SAGE, New York
27. Leary T (1957) Interpersonal diagnosis of personality: functional theory and methodology for personality evaluation. Ronald Press, New York
28. Limbrecht-Ecklundt K, Scheck A, Jerg-Bretzke L, Walter S, Hoffmann H, Traue HC (2013) The effect of forced choice on facial emotion recognition: a comparison to open verbal classification of emotion labels. Psycho-soc Med 10:1–8
29. Luciew D, Mulkern J, Punako R (2011) Finding the truth: interview and interrogation training simulations. In: The interservice/industry training, simulation & education conference (I/ITSEC), no. 1, NTSA, 2011
30. Manning CD, Schutze H (1999) Foundations of statistical natural language processing. The MIT Press, Cambridge
31. Mehrabian A (2009) Nonverbal communication. Transaction, Piscataway
32. Neff M, Wang Y, Abbott R, Walker M (2010) Evaluating the effect of gesture and language on personality perception in conversational agents. In: Intelligent virtual agents. Springer, Berlin, pp 222–235
33. Novielli N, Gentile E (2009) Modeling user interpersonal stances in affective dialogues with an ECA. In: Proceedings of the twenty-first international conference on software engineering and knowledge engineering (SEKE), pp 581–586
34. Olsen DE (1997) Interview and interrogation training using a computer-simulated subject. In: Interservice/industry training, simulation and education conference, 1997

35. op den Akker R, Bruijnes M, Peters R, Krikke T (2013) Interpersonal stance in police interviews: content analysis. Comput Linguist Neth J 3:193–216
36. Piman S, Talib AZ (2012) An intelligent instructional tool for puppeteering in virtual shadow puppet play. In: Intelligent technologies for interactive entertainment. Springer, Berlin, pp 113–122
37. Ravenet B, Ochs M, Pelachaud C (2012) A computational model of social attitude effects on the nonverbal behavior for a relational agent. In: Proceedings of workshop affect compagnon artificiel interaction (WACAI 2012)
38. Rouckhout D, Schacht R (2000) Ontwikkeling van een nederlandstalig interpersoonlijk circumplex. Diagnostiekwijzer 4:96–118
39. Scherer KR (2005) What are emotions and how can they be measured? Soc Sci Inf 44:695–729
40. Smith-Hanen SS (1977) Effects of nonverbal behaviors on judged levels of counselor warmth and empathy. J Couns Psychol 24(2):87–91
41. Spitters S, Sanders M, op den Akker R, Bruijnes M (2013) The recognition of acted interpersonal stance in police interrogations. In: 2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom), IEEE, pp 65–70
42. Strömwall LA, Granhag PA, Hartwig M (2004) 10 practitioners' beliefs about deception. In: Granhag PA, Strömwall LA (eds) The detection of deception in forensic contexts. Cambridge University Press, pp 229–250
43. Susilo AP, van Eertwegh V, van Dalen J, Scherpbier A (2013) Leary's rose to improve negotiation skills among health professionals: experiences from a southeast Asian culture. Educ Health 26:54–59
44. Traum D (2012) Non-cooperative and deceptive virtual agents. IEEE Intell Sys Trends Controv Comput Decept Noncoop 27(6):66–69
45. Vaassen F, Daelemans W (2010) Emotion classification in a serious game for training communication skills. In: Computational linguistics in the Netherlands 2010: selected papers from the 20th CLIN meeting. LOT
46. Vaassen F, Daelemans W (2011) Automatic emotion classification for interpersonal communication. In: Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, pp 104–110
47. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal processing: survey of an emerging domain. Image Vis Comput 27(12):1743–1759
48. Wauters J, Van Broeckhoven F, Van Overveldt M, Eneman K, Vaassen F, Daelemans W (2011) Delearyous: an interactive application for interpersonal communication training. In: Proceedings of CCIS serious games: the challenge
49. Wiggins JS (2003) Paradigms of personality assessment. Guilford Press, New York
50. Wilting J, Krahmer E, Swerts M (2006) Real vs. acted emotional speech. In: INTERSPEECH 2006: 9th international conference on spoken language processing, vol 2, pp 805–808